Monroe
Definitions of the terminology
of educational measurements

EDUCATIONAL RESEARCH CIRCULAR NO. 13

BUREAU OF EDUCATIONAL RESEARCH
COLLEGE OF EDUCATION

DEFINITIONS OF THE TERMINOLOGY OF EDUCATIONAL MEASUREMENTS

by

WALTER S. MONROE
DIRECTOR

PUBLISHED BY THE UNIVERSITY OF ILLINOIS

URBANA
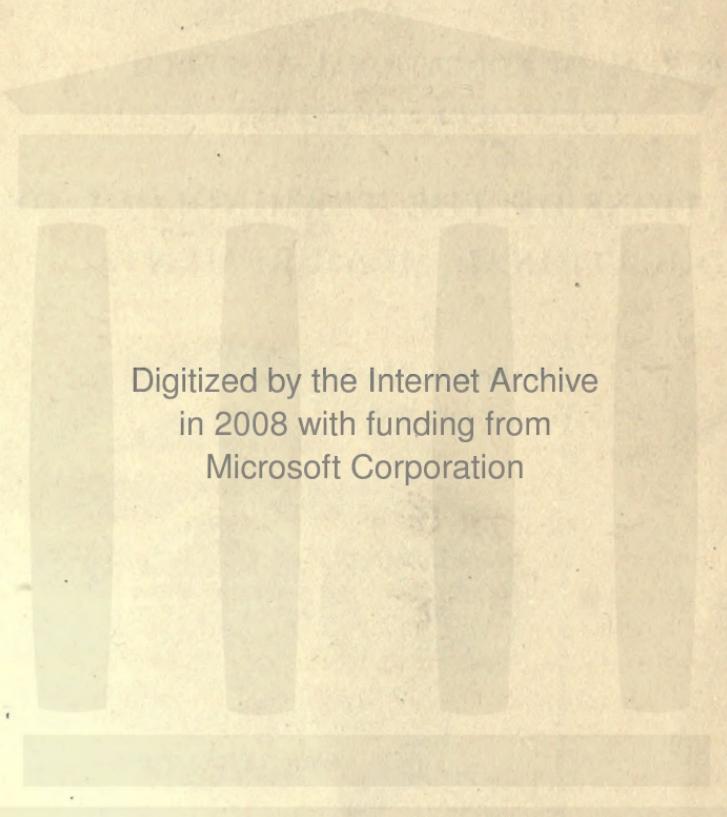
Definitions of the Terminology of Educational Measurements

Accomplishment quotient. (See Achievement quotient.)

Accuracy. (See Quality.)

Achievement age. A pupil's age score on an achievement test is frequently referred to as his "achievement age." It is simply the age which he has attained in his achievement. The field of this achievement may be limited to a particular subject in which case a pupil's achievement age is sometimes called his "subject age" to indicate the fact that the measure refers only to his achievement in a particular school subject. In this connection "educational age" has been used to denote the average of a pupil's achievements in a group of subjects which may be considered representative of his school progress.

Age norms. For calculating age norms the pupils are grouped according to age. Both chronological age and mental age have been used for this purpose. Theoretically, we should obtain the same numerical results for both groupings when unselected groups of children are used since the average mental age of a chronological age group is numerically identical with the average chronological age. Unless it is otherwise stated an age norm is the median or average of scores made by pupils ranging from the designated age up to the next. Thus the norm for 9 years is for children whose ages are between 9 and 10 years.

Age score. Age norms are used as a basis for translating point scores into age scores. For example, if the age norm for eleven years is 43 a pupil who makes a point score of 43 is said to have an age score of eleven years. Thus a pupil's age score is always interpreted as meaning that his score on the test is equivalent to the norm for the age designated by the age score. (See Age norms.)

Attainment age. (Same as Achievement age.)

Average. The average of several quantities is their sum divided by their number. When we are dealing with relatively few quantities this definition furnishes us a statement of the procedure in cal-

culating the average When we are dealing with a large number of quantities and they are grouped in a frequency distribution, the short method of calculation greatly reduces the labor required. However, the average has essentially the same meaning as when calculated by the original method.

Coefficient of correlation. The coefficient of correlation is a statistical device used to express a summary of the relationship which exists between two sets of facts that are paired together. Perfect correlation which is represented by a coefficient of 1.00 means that the two sets of facts are paired off so that the largest in one set is paired with the largest in the other, the next largest are also paired together, and so on for all pairs. Perfect negative or inverse correlation is represented by a coefficient of —1.00 which means that the largest quantity in one set is paired with the smallest in the other, the next to the largest in the first set is paired with the next to the smallest in the second, and so on. A coefficient of correlation of zero means that no relationship exists between the two sets of facts.

Coefficient of reliability. The coefficient of reliability is simply the coefficient of correlation between two sets of scores secured from two applications of the same test or from duplicate forms of it. These two applications should be separated by a relatively short time interval. For most of our educational tests the coefficient of reliability, when based upon the scores made by pupils belonging to the same school grade, ranges from .65 to .90. For a few tests coefficients of reliability .95 or higher have been reported.

Combined dimensions. Instead of describing each characteristic of a pupil's performance separately the directions for scoring some test papers provide combining the description of two or, in a few cases, three of the dimensions in a single score. For example, when the number of exercises done correctly is taken as the pupil's score on a uniform test, we have a combination of rate an daccuracy. If a scaled test is timed and the number of exercises done correctly is taken as the pupil's score we have a combination of rate, quality and difficulty.

Composite score. A composite score is the average of the scores yielded by several tests in the same field after they have been expressed in terms of a common unit and from a common zero

4

point. If the scores are averaged before this reduction is made the resulting combination will frequently be lacking in meaning because different units and different zero points are used by the different tests.

Constant error. A constant error is one which is the same for all members of a given group. This group may be a single class, a school or a group of schools. On the other hand, it may be only a division of a class as, for example, a constant error might affect only the boys in a class. A constant error may be either positive or negative, the only essential characteristic being that it is the same for all members of the group concerned.

There are two kinds of constant errors—absolute and relative. An absolute constant error has the same magnitude for all members of the group regardless of the magnitude of their scores. A relative constant error maintains a constant ratio to the magnitude of the measure. Such an error would occur in measuring a linear distance of several yards if the yard stick used was half an inch too short.

Control of testing conditions. Testing conditions include all factors other than a pupil's ability which affect or determine his performance. The most important of these factors are the following: the explanation of the tests to the pupil, the time allowed for his work, the form in which the test is presented, the pupil's physical condition, his emotional status, and the effort which he makes. Testing conditions are said to be controlled when all such factors are made the same for all the pupils taking the test, or if variations occur in any of the factors their amount is known. If the resulting scores are to be compared with the norms for a test, the testing conditions secured should be those for which the norms are stated.

Criterion measure. A criterion measure is any measure which may be used as a basis for comparison in order to determine the reliability and validity of the scores yielded by a given test. Teacher's estimates of a pupil's achievement, his school grade, and the composite scores from a number of tests are among the criterion measures that have been used. Occasionally, the scores yielded by one test will be used as a criterion measure for judging the reliability or validity of a new test.

Cycle test. In cycle tests the exercises vary in difficulty but they are so arranged that the variations occur in cycles. For ex-

ample, in a cycle test the 1st, 5th, 9th, 13th, etc., exercises might be equivalent in difficulty. The 2nd, 6th, 10th, 14th, etc., exercises would also be equivalent in difficulty. A similar condition would exist for the 3rd, 7th, 11th, and 15th exercises, and for the 4th, 8th, 12th, and 16th exercises. However, the consecutive exercises might vary widely in difficulty. A cycle of difficulty would be formed by each group of four exercises. A cycle test is useful when it is desirable to include within a single test exercises on several levels of difficulty. When such a test includes several cycles it is possible to treat it as a uniform test both in its administration and its scoring without introducing a serious error.

Derived score. Except by chance, no two tests yield point scores expressed in terms of the same unit or from the same zero point. Several proposals have been made for the calculation of a derived score which describes a pupil's performance in terms of a unit that is constant for all tests or at least for large groups of tests. Usually a point score is first obtained and this is translated into the derived score. (See Age score, Percentile score, and Quotient score.)

Diagnostic test. A diagnostic test is one which yields detailed information concerning a pupil's achievement in one or more relatively narrow fields. Frequently this type of measuring instrument consists of a number of sub-tests which yield separate measures of the pupil's achievement for a variety of fields. Such a diagnostic test can be transformed into a survey test by devising some procedure for combining the scores yielded by the separate sub-tests.

Difficulty. Difficulty has been defined as that characteristic of an exercise which when present in a large degree causes a large percent of incorrect responses and when present in a small degree is accompanied by a small percent of incorrect responses. In other words the degree of difficulty of an exercise is determined by the percent of incorrect responses obtained when it is given to a large number of pupils. If certain assumptions are made concerning the distribution of the ability of the group of pupils to whom an exercise is given and the point of zero difficulty is located, the degree of difficulty of the exercise can be expressed in terms of a measure of the variability of this distribution of ability. This unit is the difference in difficulty between two exercises which are answered correctly by a certain percent of a given group of pupils. The median devia-

tion (P. E.) is frequently used as a unit. It is defined as the difference in difficulty between an exercise which is answered correctly by 50 percent of the pupils and an exercise which is answered correctly by only 25 percent of the same pupils. The standard deviation (S. D. or σ) is also used as a unit. It is the difference between an exercise answered correctly by 50 percent of the pupils and an exercise answered correctly by only 15.87 percent of the same pupils. Thus we may describe the difficulty of exercises as being 2.7 P. E., 6.3 P. E., 5.2 σ, etc.

Difficulty score. A difficulty score is a statement of the highest level of difficulty on which a pupil has done the exercises with a specified or standard degree of accuracy. This score is yielded only by scaled tests.

Dimensions of a pupil's performance. A pupil's performance is described in terms of its distinguishing characteristics. These are (1) its amount or when produced under timed conditions, the rate of work, (2) the quality or accuracy of the performance and (3) the level of difficulty upon which it was given. These three characteristics are sometimes spoken of as the dimensions of the pupil's performance. (See Rate, Score, Quality, Difficulty, and Combined dimensions.)

Discrimination. A test is said to be lacking in discrimination when it fails to give different scores to pupils who are known to differ in ability. This may happen to only a few of the pupils to whom the test is given. For example, a very easy test lacks discrimination for those pupils who make perfect scores. A very hard test is lacking in discrimination for those who make zero scores. A lack of discrimination may be indicated by other evidence. If a distribution of scores differs conspicuously from the normal distribution, when we have reason to believe that the distribution of true scores would approximate the normal, we have evidence of lack of discrimination for certain pupils. If two groups are known to differ in ability, as for example, a fifth grade group and a sixth grade group, a test which fails to yield a higher average score for the sixth grade group than for the fifth grade group is lacking in discrimination. There will also be a lack of discrimination for certain pupils if the unit used is so large that pupils who differ in ability receive identical scores.

Educational objectives, agreement with. In selecting exercises for the final form of a test they may be examined with reference to their agreement with certain educational objectives. For example, in constructing his spelling scale Ayres selected certain words on the basis of their frequency of use in adult writing. Charters selected exercises for his language and grammar tests which are in agreement with the language errors made by children. In the case of other tests the consensus of opinion of competent persons has been used as a guide in the selection of exercises. (See also Statistical selection.)

Exercises. The exercise is a structural unit of a test. Some of the simpler types call for a word to be spelled, an example to be worked, or a question to be answered. Other exercises are more complex. Some are large, in that they consist of several items and require much time for completion. A test usually consists of a considerable number of exercises, but occasionally of a single long exercise.

Fore exercise. A fore exercise is a preliminary test which has for its purpose, acquainting a pupil with the character of the exercises which he is asked to do in the test. The pupil's performance on the fore exercise is not included in computing his score.

Form. The term "form" is practically always used in the sense of a duplicate form. Thus a test is said to have more than one form when there are duplicate measuring instruments consisting of similar but not of identical exercises. Such duplicate forms are intended to yield equivalent measures. Hence, when the two forms are administered under exactly the same conditions, a pupil should make the same score on one form that he makes on another. Investigation has shown that, in general, duplicate forms do not yield equivalent measures even when a great deal of care has been exercised in their construction. Hence, when making comparisons between scores yielded by duplicate forms, it is necessary to know concerning their degree of equivalence and to make corrections for any differences which may have been ascertained.

The "form" of a test should be distinguished from "parts" and "division." In a few cases "part" has been used with a meaning very similar to "exercise" but it is generally used to designate a section or division of the measuring instrument which is designed for certain grades. This use is illustrated by Part 1 and Part 2 of

Thorndike's Scale for the Understanding of Sentences. "Division" usually has the same meaning. In a few cases "part" has been used with a different meaning. In some cases a test has been divided into "parts" without the term being used. For example, Monroe's Standardized Silent Reading Tests consist of three parts or divisions although neither of these terms has been used in connection with its title. Test I is designed for grades 3, 4, and 5, Test II for grades 6, 7, and 8, and Test III for the high school. When a measuring instrument has parts or divisions (not sub-tests) the total instrument more probably would be described as a series or a group of instruments with different parts or divisions which are designed to measure the ability of pupils on different levels.

Function. The function of a test is a statement of the ability which it is designed to measure plus a statement of the type of information which it will yield concerning this ability. A pupil's performance is completely described in terms of three dimensions. The score which a given test yields may be restricted to a single dimension or it may involve two or even three, separately or in combination. A statement of the function of the test should also include some specification of its scope. A test may be very general in scope, in which case it is called a general or survey test. If it yields measures for relatively narrow fields it is called a detailed or diagnostic test. Certain tests have a prognostic function.

Grade norms. Grade norms are the averages or medians of the scores made by pupils in the respective school grades. In some cases a grade refers to an entire year's work. In other cases it represents only a semester's work. Usually when grade norms are stated it is understood that there are eight years in the elementary school and four years in a high school. When such norms are applied to a system which has seven or nine years below the high school, it is necessary to make adjustments.

Index of reliability. The index of reliability differs from the coefficient of reliability in that it is the coefficient of correlation between a set of obtained scores and the corresponding set of true scores rather than the coefficient of correlation between two sets of obtained scores. It is calculated from the coefficient of reliability by the following formula in which r_{12} represents the coefficient of reliability and r_{1t} the index of reliability.

$$r_{1t} = \sqrt{r_{12}}.$$

Irregular test. An irregular test is one in which the exercises vary in difficulty and are not arranged in order of ascending or descending difficulty. Irregular tests usually result when exercises are selected on some basis other than that of difficulty. When extreme irregularities are avoided irregular tests may be treated as uniform tests without introducing serious errors.

Median. The median of a set of scores, arranged in ascending or descending order of magnitude, is the middle score, or when there is no middle score it is the average of the two middlemost scores.

Mental age. A pupil's age score on an intelligence test is called his mental age.

Normal distribution. A normal distribution is symmetrical. At either extreme there are very few measures. Most of the measures are grouped near the center and there is a rather gradual decrease down to zero at the extremes. Distributions which approximate a true normal distribution are generally described as normal distributions.

Norms. The norms for an educational test are determined by having the test given to a large number of pupils belonging to several groups and by taking the average or median of these scores. Thus our present norms are the average or median achievements of pupils. In most of our uses of norms we have assumed that the average or median of present achievement is that which the pupils should achieve. It has been suggested that "standard" be used to designate the scores which pupils *should make* thereby making a distinction between "norm" and "standard" but our common practice is to use the two terms with the same meaning. A test for which norms have been determined is said to be standardized. Norms may be obtained for both grade groups and age groups. (See Age norms and Grade norms.)

Objective. A measuring instrument is said to be objective when different persons using it to measure the same thing secure approximately the same result. The opposite of objective is subjective. Both of these terms are relative. No educational tests are absolutely objective but those which are rather highly objective are commonly spoken of as objective tests. The scoring of a test is said to be objective when different scorers will in general assign the same scores to the same papers. (See Subjective.)

Overlapping. The term "overlapping" is used to describe the relative position of two distributions. Its most frequent use is in the case of distributions for successive grade groups or successive age groups. The percent of one distribution which is beyond the median or average of the other distribution is taken as the measure of the overlapping.

Percentile scores. A percentile score describes the pupil's place in the distribution of the scores of the group to which he belongs. Consider, for example, the distribution of scores of a large number of fifth grade pupils. Locate a pupil's score on the base line of the distribution. The position of this point can be described by telling the percent of the total scores in the distribution which are below his score. For example, if 82 percent of the scores are below his, he may be said to have an 82 percentile score. If a standard distribution has been secured tables may be prepared by means of which it is relatively easy to translate any point score into the corresponding percentile score.

Performance. A pupil's performance is what he does. The performance is usually written and for testing purposes must be such that it can be easily observed by any competent observer. A performance is sometimes described as objective which means that the result, when observed by different persons, is the same.

Point score. A point score is the score which is yielded directly by the test. Exercises done correctly, the number of exercises attempted, and the level of difficulty reached are point scores. The magnitude of a point score depends upon the size of the unit which is usually determined by the exercises, and the length of the test. It is only by chance that two tests yield point scores in terms of the same unit and expressed from the same zero point. (See Derived score.)

Power test. The term "power test" is most frequently used to describe a scaled test which yields only a difficulty score. Such a measuring instrument has been called a power test since it measures the power or ability of the pupils to do increasingly difficult exercises of the same kind. With only a slight change in the meaning other types of tests could be called power tests when only the accuracy or quality score is used. A power test is not timed.

Practice effect. Practice effect refers to the average increase of the scores of one trial over those yielded by a preceding trial, when there has been no opportunity for coaching between the two administrations of the test. Because of becoming acquainted with the nature of the exercises pupils tend to make higher scores on the second trial of a test than they did on the first. This practice effect constitutes a constant error when the same norms are used to interpret the scores from both trials. The magnitude of this error varies with different tests but in general second-trial scores are on the average ten percent greater than first-trial scores.

Preliminary test. (Same as Fore exercise.)

Probable error of estimate. The probable error of estimate is a statistical device derived from the coefficient of correlation which is helpful in interpreting cases of "high" correlation. It may be defined as the measure of departure from the perfect correlation. This is given in terms of the median deviation or P. E. of the distribution of all the departures from perfect correlation in the pairs of scores from which the coefficient of correlation was calculated. It is calculated from the coefficient of correlation by the following formula in which $P. E._{Est}$ designates the probable error of estimate, σ_2 is the standard deviation of the distribution of scores obtained from the second application of the test, and r_{12} is the coefficient of correlation between two sets of obtained scores.

$$P. E._{Est} = .6745 \ \sigma_2 \sqrt{1 - r_{12}^2}$$

A probable error of estimate of 3.4 means that in 50 percent of the pairs of scores there is a departure of the second score from a perfect correlation with the first score of more than 3.4 in 50 percent of the pairs.

Probable error of measurement. The probable error of measurement bears the same relation to the probable error of estimate that the index of reliability bears to the coefficient of reliability. In other words it is a measure of the departure of a given set of obtained scores from perfect correlation with the corresponding true scores. It is calculated from the coefficient of reliability by the following formula in which $P. E._M$ is the probable error of measurement, σ is the average of σ_1 and σ_2 and r_{12} is the coefficient of correlation between two sets of obtained scores.

$$P. E._M = .6745 \ \sigma \ \sqrt{1 - r_{12}}.$$

12

A probable error of measurement of 5 means that in 50 percent of the cases the obtained score will differ by as much or more than 5 from the pupil's true score. In 50 percent of the cases the difference will be less.

Prognostic test. A prognostic test is a test which has for its function the prediction of a pupil's status at some future time. This prediction, of course, is based upon the pupil's performance at the present time. All tests have some prognostic value, but certain tests which have been devised with special reference to this function are called prognostic tests.

Quality. The quality of a pupil's performance is sometimes described in terms of the percent of the exercises which he has done correctly. In such cases quality is synonymous with accuracy. Certain types of performances (for example, a specimen of handwriting) cannot be classified as right or wrong. In such cases quality means merit and it is described in terms of a quality scale.

Quotient score. A point score or an age score is simply a description of the absolute amount of a pupil's achievement or general intelligence. Such absolute measures are significant only when compared with appropriate norms. For this reason it has been proposed to divide the point scores or age scores by certain other measures of the pupil. For example, a pupil's mental age divided by his chronological age gives a quotient which is called the intelligence quotient or I. Q. A pupil's achievement age divided by his mental age gives the achievement quotient or A. Q. More strictly speaking the A. Q. is the quotient of a pupil's achievement age divided by the norm for his mental age. Other quotients have been proposed. For example, a pupil's achievement age divided by his chronological age gives the educational quotient or E. Q. The educational quotient divided by the intelligence quotient has been called the accomplishment quotient or A. Q. This, however, is identical with the achievement quotient described above.

Rate score. A rate score is a measure of a pupil's rate of work. It is usually expressed in terms of the number of exercises or the number of units of work which he has attempted within a given time limit. It may, however, be expressed as the number of minutes or seconds used by a pupil to complete a specified amount of work.

Rate test. A rate test is one which yields a rate score. It may yield other scores also but it is essential that it yields a rate score which is unaffected by the other dimensions of the pupil's performance.

Reliability. The reliability of a test describes the extent to which a second application of a test will yield scores equivalent to the first. It is a well known fact that when a test is administered the second time some pupils will make higher scores and some lower. These changes are due, for the most part, to the presence of variable errors in both sets of scores. The reliability of a test is the description of the magnitude of these variable errors. Any constant errors produced by practice effect or by inaccurate timing or by other conditions which effect the entire group are not included in the reliability. (See Coefficient of reliability, Index of reliability, Probable error of estimate, and Probable error of measurement.)

Scale. When used in a restricted sense the word "scale" designates that portion of a measuring instrument which is used in describing a pupil's performance. In the case of some of our measuring instruments the scale is conspicuous, as for example, in Willing's Scale for Measuring Written Composition. This scale is used only in describing the performance of pupils. In order to secure a suitable performance it is necessary to follow certain directions which are not, strictly speaking, a part of this scale. In other measuring instruments, such as Courtis Standard Research Tests in Arithmetic, Series B, the scale is less obvious. There is, however, in every measuring instrument a scale which functions in the description of the performances secured from the pupils. The word "scale" is used also in a general sense to designate the total measuring instrument. Usually this is done only when the scale for describing the pupil's performance is the distinguishing characteristic of the measuring instrument. (See Test.)

Scaled test. A scaled test is one in which the exercises are arranged in order of ascending difficulty. Usually, the increase in difficulty from one exercise to the next is approximately constant throughout the scale. This is a desirable but not necessary feature. Another essential characteristic of the scaled test is that the exercises of least difficulty be sufficiently easy so that all pupils to whom the test is given will be able to do them and that the most difficult

exercises be such that practically no pupils will be able to do them correctly.

Score. A pupil's score is a description of his performance. There are several types of scores, each of which has its own function. (See Rate score, Accuracy, Quality, Difficulty, Point score, Derived score, Combined dimensions.)

Selection of exercises. Usually in constructing educational tests a large number of exercises are secured and from this collection those to be used in the final test are selected. There are three criteria of selection which are frequently used, sometimes singly and sometimes in combination: (1) statistical selection, (2) agreement with educational objectives, and (3) suitableness for testing purposes as determined by trial. Occasionally the selection is made by the author of the test without the guidance of definite criteria. Such selection may be described as arbitrary. (See Statistical selection and Educational objectives.)

Spiral test. The word "spiral" has been used to describe a measuring instrument which consists of several sub-tests so arranged that in general there is an increase in difficulty in the successive sub-tests. A good example of this type of test is the Cleveland Survey Arithmetic Test.

Standards. (See Norms.)

Standardized test. A test is said to be standardized when norms or standards have been determined for it. The standardization of the test has no reference to the selection of the exercises or to the unit in terms of which the point score is expressed. In the field of physical measurement the standardization of a measuring instrument has a different meaning. It refers to the fixing of the magnitude of the unit. For example, the standardization of linear measures means fixing the precise length of the fundamental unit —the yard. This meaning of standardization is approached in some of the proposed derived scores.

Statistical selection of exercises. The usual procedure in constructing an educational test is to secure a rather large collection of exercises. From this list certain exercises are selected. One method for making this selection is to ascertain the percent of correct responses for each exercise and from this to compute their diffi-

culty. Those exercises are then selected whose degree of difficulty is appropriate for the structure of the desired test. Such a selection is said to be statistical. (See Educational objectives.)

Subjective. An educational test is said to be subjective when different persons or the same person at different times, using it to measure the same thing, secure different results. The source of the subjectivity may be in the giving of the tests to the pupils or in the scoring of the test papers. In the latter case the scoring or the description of the pupil's performance is said to be subjective. This means that different persons will tend to assign different scores to the same papers. It should be noted that "subjective" and "objective" are relative terms. All educational tests are subjective in some degree. Certain tests are very highly subjective and others are only very slightly so. As the term is generally used a subjective test is one which is highly subjective. (See Objective.)

Sub-test. Some measuring instruments consist of major divisions which are called sub-tests. For example, the Cleveland Survey Test in Arithmetic is a measuring instrument which consists of fifteen sub-tests. Each sub-test is made up of a number of exercises. (See Exercise.)

Survey test. A survey test is one which is general in its scope. It is usually made up of a number of sub-tests covering a variety of fields of subject-matter. The scores yielded by these sub-tests may or may not be combined into a single score. The function of a survey test is to yield a general or average measure of a pupil's achievement over a large field. Sometimes this field may be restricted to certain divisions within a subject as, for example, arithmetic, or it may include several school subjects.

Test. The word "test" is used both in a general sense and in a restricted sense. In the general sense it is used to designate any type of instrument for measuring mental ability. Thus it may be used in referring both to instruments which have been named "tests" and to instruments which have been named "scales" by their authors. In the restricted sense it refers to the portion of a measuring instrument that is used to secure a performance from the pupil. Some of our measuring instruments are spoken of as tests and others as scales but there is little evidence of discrimination in the use of these terms. In so far as there has been discrimination in respect to "test" and

"scale" that term has been used which was most characteristic of the distinguishing feature of the measuring instrument. For example, we have the Courtis Arithmetic Tests, the Kansas Silent Reading Test, and the Thorndike Handwriting Scale. (See Scale, Uniform test, Scaled test, Irregular test, Cycle test, and Spiral test.)

Time limit. A test is said to be "timed" when the time allowed is such that a measure of the rate of work of the pupils can be secured. Usually this means that the time limit is such that practically no pupils will be able to finish the test. All types of test may be timed but the time limit is most significant in the case of a uniform test. When applied to a scaled test, if the time limit is such that practically all pupils are able to advance as far along the scale as their ability permits before time is called, the test is essentially untimed. Although a time limit may be specified in such a case it is not incorrect to say that the pupils are allowed practically unlimited time or all the time they need.

True score. A pupil's true score is defined as the average of a large number of measurements of a given ability made under the same conditions. It is, of course, impossible to make even a second measurement of a pupil's ability under exactly the same conditions as the first measurement was made because the taking of the test in itself has changed one factor of the testing conditions. For this reason it is impossible to obtain a true score by averaging the scores obtained from the repeated applications of a test. However, the concept of a true score is frequently helpful and we are able to make certain statistical calculations with reference to true scores even though it is impossible to obtain them. (See index of reliability and Probable error of measurement.)

Uniform test. A uniform test is one whose exercises are approximately equivalent in difficulty. Generally the exercises are also similar in content. This equivalence in difficulty may be secured by constructing exercises of the same sort as, for example, in the Courtis Standard Research Tests in Arithmetic, Series B, or by selection on a statistical basis.

Validity. The term "validity" refers to the truthfulness with which a test fulfills its function. A test may fail to do this by reason of inaccurate scores or by failing to measure the ability specified by its function. A test whose score is lacking in accuracy is said

to be unreliable. Such a test can never be highly valid. Becaus
we are not able to obtain completely valid measures for purpose
of comparison it is necessary to use certain indirect and partia
methods in determining the validity of a given test. (See Subjective
Reliability, and Discrimination.)

Variable errors. Variable errors are different for the differen
members of a group. Approximately half are positive, some are
zero and the remainder are negative. The distinguishing charac
teristic of all variable errors is this difference from pupil to pupil
Unless highly accurate measures of the same trait are available
for comparison we are not able to determine the magnitude of the
variable error for a particular pupil. The best we are able to do
is to state what the chances are that the variable error does not
exceed a certain magnitude in a particular case.

370.78 Illinois. University.
I29E   Bureau of Educational
no. 13   Research.
  Monroe   Educational research
           circular)
         Definitions of the terminology
         of educational measurements